(12) UK Patent Application (19) GB (11) 2 349 260 (13) A

(43) Date of A Publication 25.10.2000

(21) Application No 9909429.4

(22) Date of Filing 23.04.1999

(71) Applicant(s)
Canon Kabushiki Kaisha
(Incorporated in Japan)
30-2 3-chome Shimomaruko, Ohta-ku, Tokyo, Japan

(72) Inventor(s)
Yuan Shao

(74) Agent and/or Address for Service
Beresford & Co
2-5 Warwick Court, High Holborn, LONDON,
WC1R 5DJ, United Kingdom

(51) INT CL[7]
G10L 15/06 , G06K 9/66 , G10L 17/00 // G10L 15:10
15:12

(52) UK CL (Edition R)
G4R RPD RRL R1C R1F R1X
U1S S2125 S2126 S2210 S2243

(56) Documents Cited
US 4751737 A

(58) Field of Search
UK CL (Edition R) G4R RPD RPN RRL
INT CL[7] G06K 9/00 9/20 9/32 9/60 9/62 9/64 9/66 9/78
9/80 , G10L 15/00 15/06 17/00
Online:WPI, EPODOC, JAPIO

(54) Abstract Title
Training apparatus

(57) A reference model is generated from three or more training signals. The system simultaneously compares and aligns the three or more training signals with each other and, from the alignment results, generates a reference model representative of the training signals. The system preferably employs a multi-dimensional dynamic programming algorithm to perform the comparison and alignment, and the subsequent combination preferably combines the signals onto an averaged time axis.

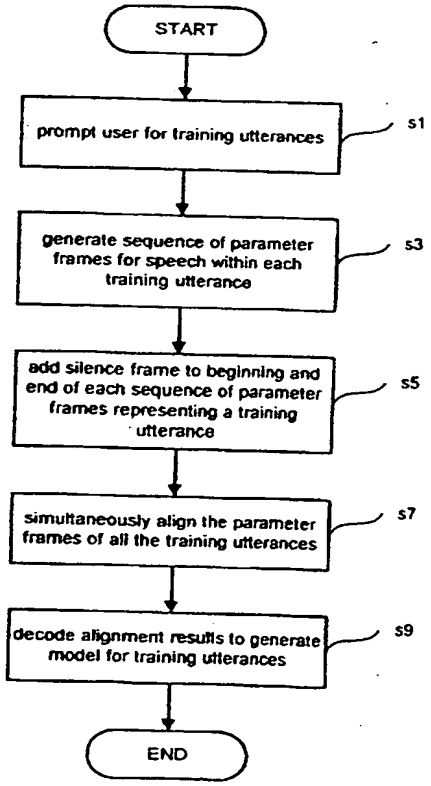


Fig. 3b

GB 2 349 260 A

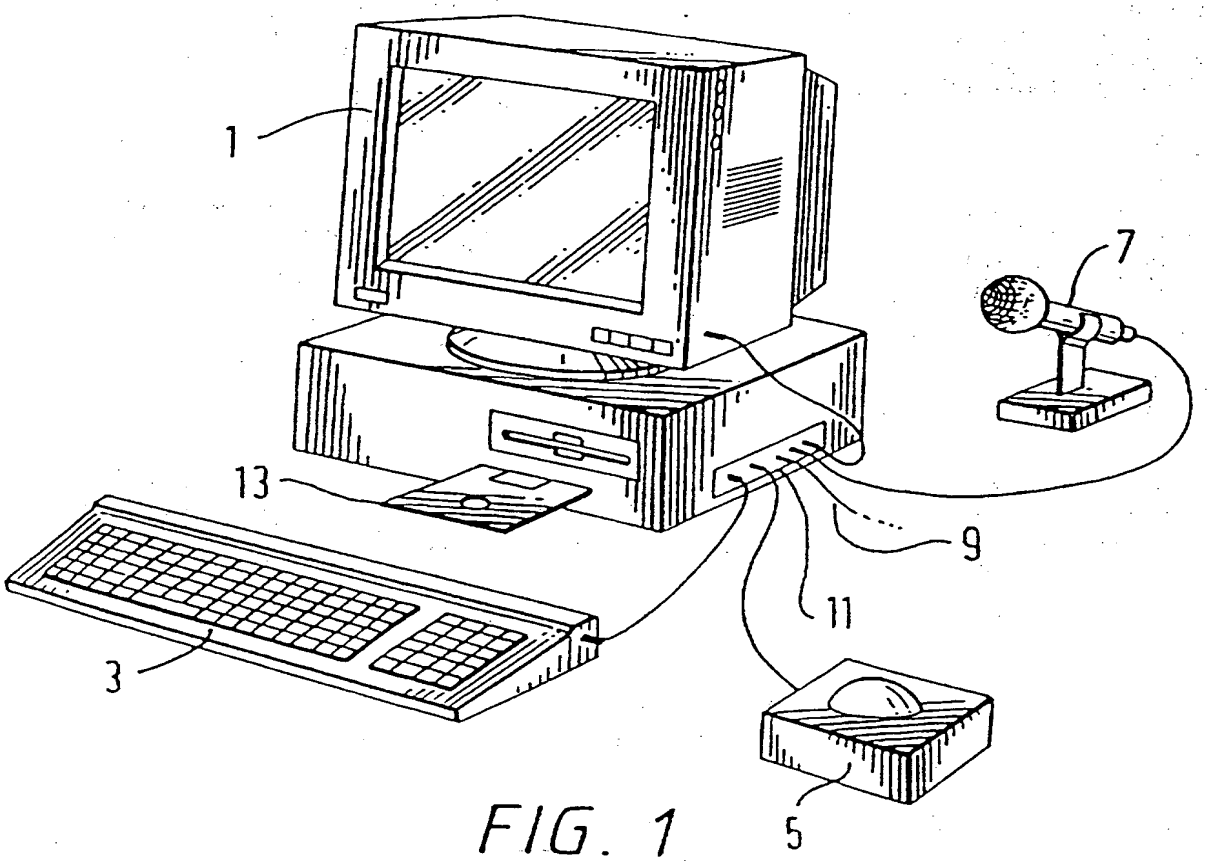
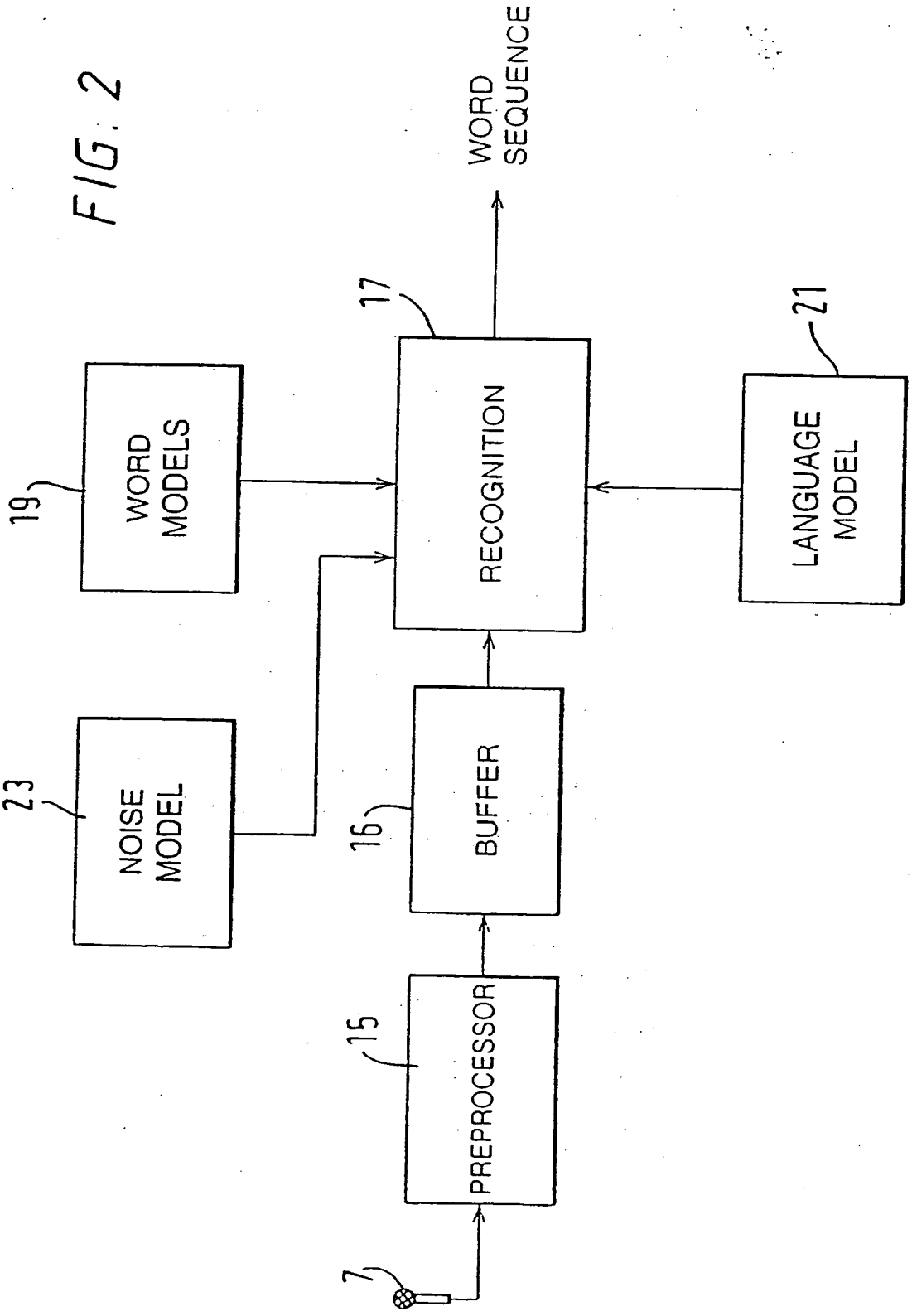


FIG. 1

FIG. 2



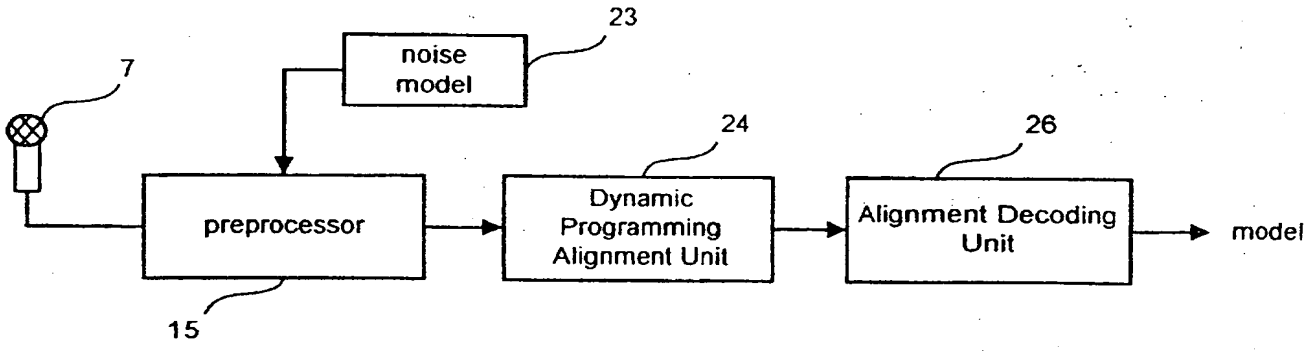


Fig. 3a

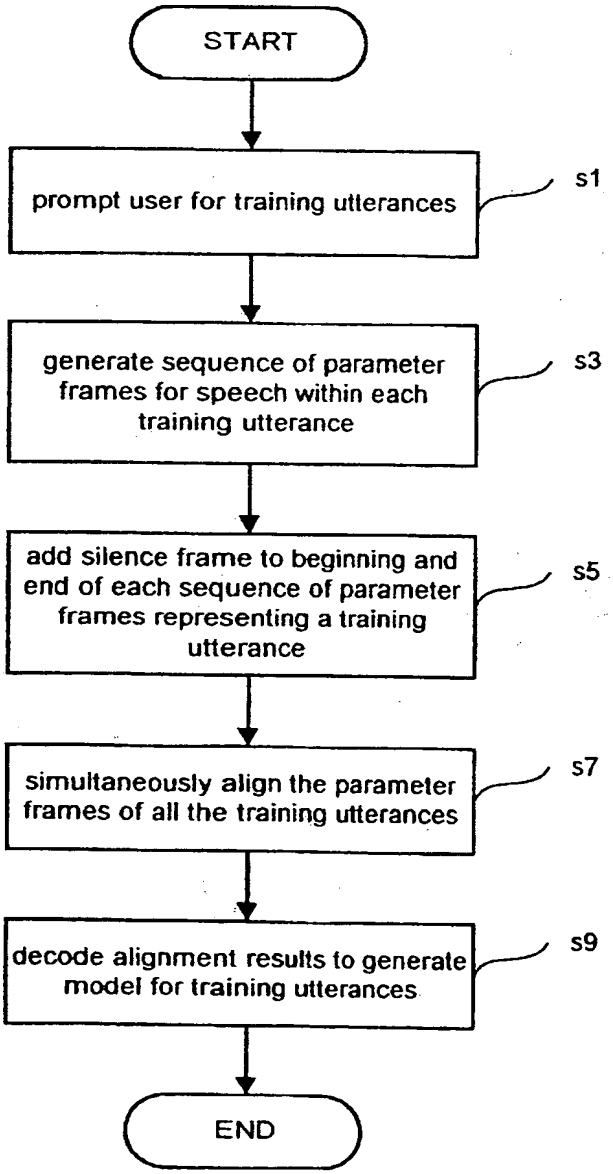
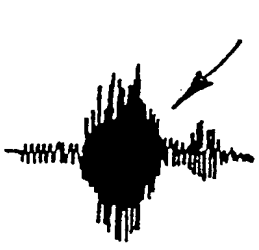


Fig. 3b

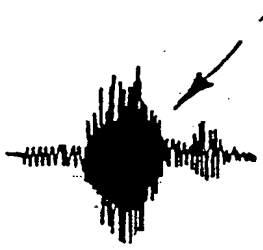
$$\left[f_0^{W1}\ f_1^{W1}\ f_2^{W1}\ f_3^{W1}\ f_4^{W1}\ f_5^{W1}\ \ldots\right]$$

31



$$\left[f_0^{W1}\ f_1^{W1}\ f_2^{W1}\ f_3^{W1}\ f_4^{W1}\ f_5^{W1}\ \ldots\right]$$

33



$$\left[f_0^{W1}\ f_1^{W1}\ f_2^{W1}\ f_3^{W1}\ f_4^{W1}\ f_5^{W1}\ \ldots\right]$$

35

FIG. 4

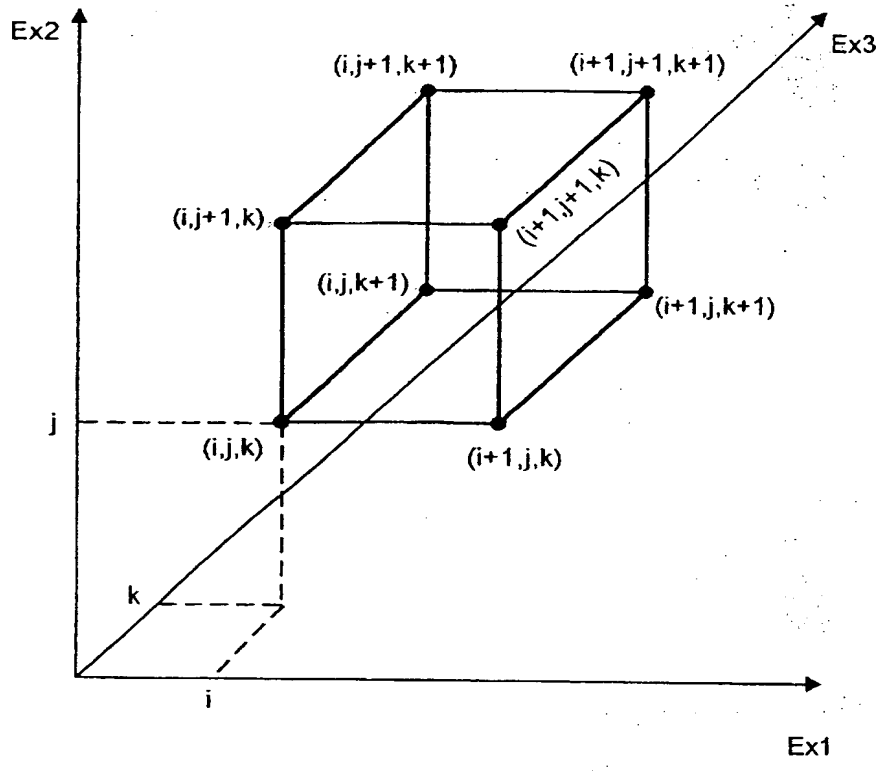


Fig. 5

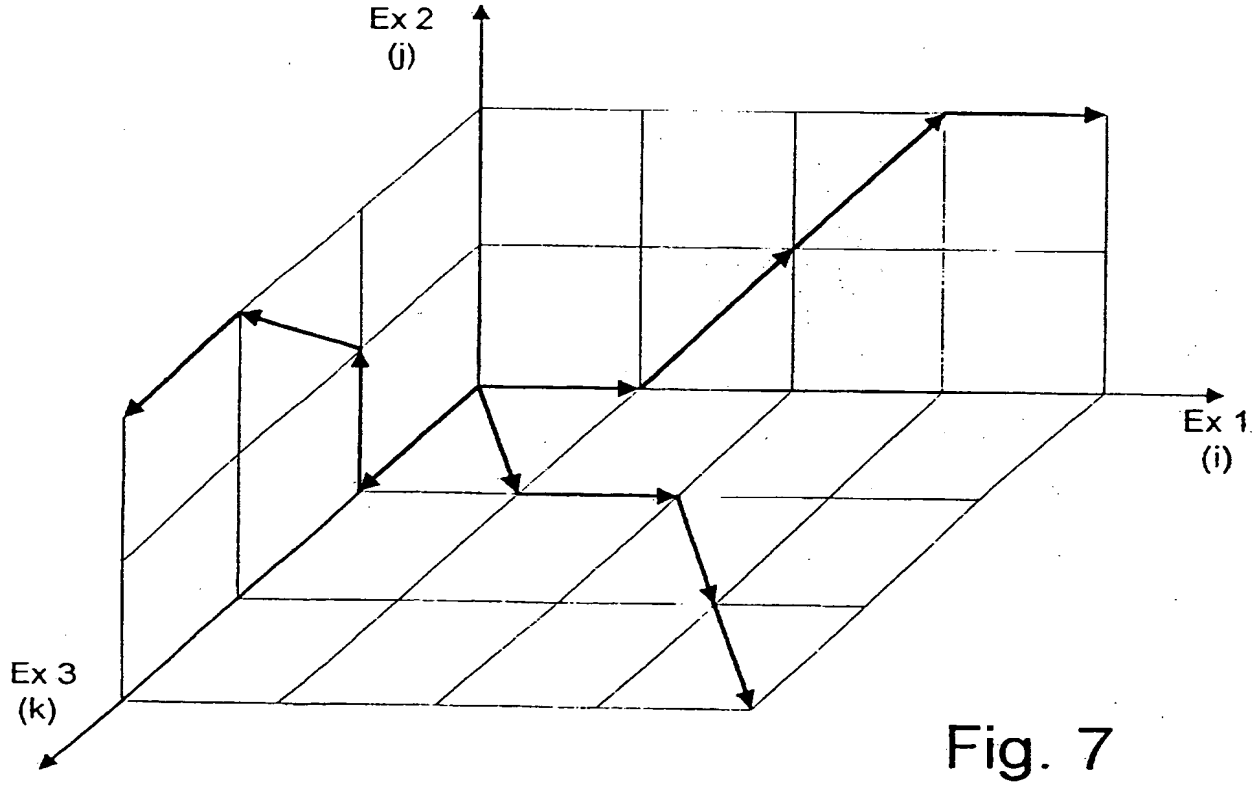


Fig. 7

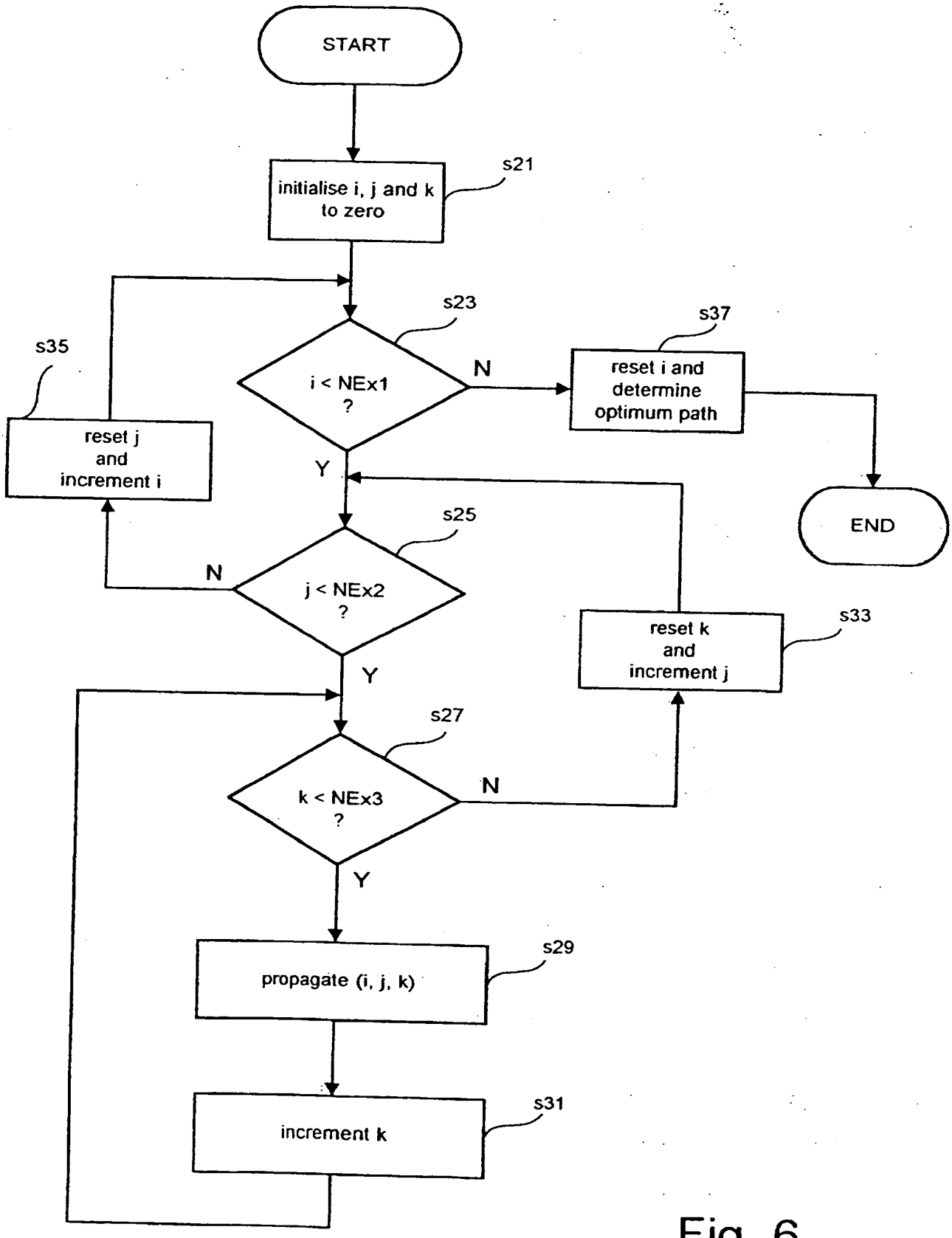


Fig. 6

TRAINING APPARATUS AND METHOD

The present invention relates to an apparatus and method for generating signal models for use in subsequent comparison processors. The application has particular although not exclusive relevance to the training of word models for use in speech recognition or speaker verification systems.

All speech recognition systems operate by receiving an input utterance from a user and comparing the utterance with stored reference models to provide a recognition result. A problem with this kind of speech recognition system is in the generation of accurate and reliable reference models.

US-4751737 describes a technique for generating models for reference words from two or more utterances of the words input by the user. The system divides the speech signal for each utterance into a sequence of time frames and determines a set of parameters for each time frame representative of the input speech signal during the time frame. It then aligns the parameter frames of the first utterance of a word with the parameter frames of a second utterance of the word using a dynamic programming alignment algorithm. The aligned frames are then combined onto an averaged time axis. In other words, if

the first utterance is represented by ten parameter frames and the second utterance by eight parameter frames, then the model which is generated will comprise nine parameter frames. If the system has more than two utterances of the word, then it takes the model which was generated from the first two utterances and aligns it with the frames of the third utterance of the word and so on, until all the utterances have been used to generate the model. This patent also teaches that the data representative of any "interim" model should be weighted by the number of input utterances which have been used to generate the interim template, so as to place more weight on the parameter frames of the interim model than the parameter frames of the new utterance.

The present invention aims to provide an alternative technique for generating a model from three or more training examples.

According to one aspect, the present invention provides an apparatus for generating a reference model representative of three or more training signals, the apparatus comprising: means for simultaneously comparing and aligning the training signals and means for generating the reference model using the alignment results. The inventors have found that by aligning all the training utterances simultaneously in this manner,

rather than two at a time as suggested by the prior art, results in approximately a 20% reduction in the number of recognition errors when the models are subsequently used during a speech recognition process.

Preferably, the apparatus employs a multidimensional dynamic programming alignment technique to align the training utterances, since it can determine the optimum alignment between the training utterances. The system also preferably combines the aligned utterances onto an average time axis, since such a model is more representative of all of the training utterances.

According to this aspect, the present invention also provides a corresponding method and computer software for programming a programmable processor to carry out the method of the present invention.

An exemplary embodiment of the present invention will now be described with reference to the accompanying drawings, in which:

Figure 1 is a schematic view of a computer which may be programmed to operate an embodiment of the present invention;

Figure 2 is a schematic overview of a speech recognition

system;

Figure 3a is a block diagram of the processing circuits employed during the training operation;

Figure 3b is a flow chart illustrating the process steps involved in generating a reference model from a plurality of training utterances;

Figure 4 illustrates three utterances of a training word and the corresponding sequence of parameter frames for each utterance of the training word generated by a preprocessor circuit which forms part of the speech recognition system shown in Figure 2;

Figure 5 is a three dimensional cartesian plot, with one axis provided for each one of three training utterances, and showing eight lattice points which represent the possible transitions which are allowed by a dynamic programming alignment operation which aligns the three training utterances;

Figure 6 is a flow chart illustrating the steps involved in performing a three dimensional dynamic programming matching operation; and

Figure 7 is a three dimensional cartesian plot

illustrating an alignment path between three training examples.

Embodiments of the present invention can be implemented in computer hardware, but the embodiment to be described is implemented in software which is run in conjunction with processing hardware such as a personal computer, workstation, photocopier, facsimile machine or the like.

Figure 1 shows a personal computer (PC) 1 which may be programmed to operate an embodiment of the present invention. A keyboard 3, a pointing device 5, a microphone 7 and a telephone line 9 are connected to the PC 1 via an interface 11. The keyboard 3 and pointing device 5 enable the system to be controlled by a user. The microphone 7 converts the acoustic speech signal of the user into an equivalent electrical signal and supplies this to the PC 1 for processing. An internal modem and speech receiving circuit (not shown) may be connected to the telephone line 9 so that the PC 1 can communicate with, for example, a remote computer or with a remote user.

The programme instructions which make the PC 1 operate in accordance with the present invention may be supplied for use with an existing PC 1 on, for example, a storage device such as a magnetic disc 13, or by downloading the

software from the internet (not shown) via the internal modem and the telephone line 9.

The operation of the speech recognition system of this embodiment will now be briefly described with reference to Figure 2. A more detailed description of the speech recognition system can be found in the Applicant's earlier European patent application EP 0789349, the content of which is hereby incorporated by reference. Electrical signals representative of the input speech from, for example, the microphone 7 are applied to a preprocessor 15 which converts the input speech signal into a sequence of parameter frames, each representing a corresponding time frame (in this embodiment 16 milliseconds) of the input speech signal. The sequence of parameter frames are supplied, via buffer 16, to a recognition block 17 where the speech is recognised by comparing the input sequence of parameter frames with reference models or word models 19, each model comprising a sequence of parameter frames expressed in the same kind of parameters as those of the input speech to be recognised.

A language model 21 and a noise model 23 are also provided as inputs to the recognition block 17 to aid in the recognition process. The noise model is representative of silence or background noise and, in

this embodiment, comprises a single parameter frame of the same type as those of the input speech signal to be recognised. The language model 21 is used to constrain the allowed sequence of words output from the recognition block 17 so as to conform with sequences of words known to the system. The word sequence output from the recognition block 17 may then be transcribed for use in, for example, a word processing package but, in this embodiment, it is used as operator commands to initiate, stop or modify the action of the PC 1.

A description will now be given with reference to Figures 3 to 5 of the way in which the word models 19 are generated for the speech recognition system shown in Figure 2.

Figure 3a is a block diagram illustrating the circuitry employed during the training operation and Figure 3b is a flow chart illustrating the processing steps involved in the training operation, which is initiated by the user via an appropriate user interface (not shown). As shown in Figure 3b, the first step s1 of the training operation prompts the user to input training utterances of the words or phrase to be modelled via the microphone 7. In this embodiment, the prompt is made via the display of the computer 1. If the training utterances are not to be associated with an existing command, then in step s1,

the user will also associate a computer response for the model which is to be generated from the training utterances. As the user is inputting the training utterances, the preprocessor 15 processes the input signal from the microphone and generates, in step s3, a sequence of parameter frames representative of the speech within each training utterance.

Figure 4 illustrates the result of this inputting stage, when the user inputs three utterances 25, 27 and 29 of the command "copy". As shown in Figure 4, the preprocessor generates a respective sequence of parameter frames 31, 33 and 35 representative of the corresponding utterances. Once the parameter frames have been generated for the training utterances, the preprocessor 15 adds, in step s5, the noise or silence frame (i.e. the noise model 23 shown in Figure 2) to the beginning and end of each sequence of parameter frames (the reason for which will be described later). Then, in step s7, the dynamic programming alignment unit 24 simultaneously compares and aligns the sequences of parameter frames 31, 33 and 35 of each of the training utterances with each other. In this embodiment, this alignment operation is performed using a multi-dimensional dynamic programming alignment algorithm. After the parameter frames for each training utterance have been aligned with the parameter frames of the other utterances, the alignment decoding

unit 26 decodes, in step s9, the alignment results to generate a model for the training utterances. The inventors have found that by aligning all the training utterances simultaneously in this manner, rather than two at a time as suggested in the prior art, results in a 20% reduction in the number of recognition errors, when the models are subsequently used in a speech recognition system. This is because the model which is generated does not depend upon the order in which the training utterances are combined, as it does using the prior art technique of combining two models at a time.

As mentioned above, the noise model 23 was added to the beginning and end of each of the training utterances. The reason for this will now be explained. The parameter frames generated by the preprocessor for each of the training utterances are likely to include one or more parameter frames at the beginning and end thereof which correspond to background noise or silence. If the noise model 23 is not added to both ends of the training utterances, then the alignment step s7 and the decoding step s9 will not be able to identify that these frames correspond to background noise, and therefore, the resulting model will inevitably include frames which correspond to noise rather than the speech command. However, when the noise model 23 is added to the beginning and end of each of the training utterances, the

parameter frames in the training utterances which correspond to noise can be identified and therefore disregarded during the decoding step s9, because they should align with the noise model 23 rather than any frames in the other training utterances which also correspond to noise. This is because the noise model 23 represents an average of a number of background noise frames and therefore, on average, the variation between the parameter frames in the training utterances which correspond to noise and the noise model 23 should be less than the variation between the parameter frames corresponding to noise in one of the utterances and those which correspond to noise in the other utterances.

A more detailed description of the alignment step s7 and of the decoding step s9 will now be given with reference to Figures 5 to 7.

ALIGNMENT

As those skilled in the art will know, dynamic programming is a technique which can be used to find the optimum alignment between the sequences of parameter frames representative of the training utterances. It does this by simultaneously propagating a plurality of dynamic programming paths, each of which represents a possible matching between a sequence of parameter frames from each of the training utterances. In order to

determine the optimum alignment between the sequences of parameter frames representative of the training utterances, the dynamic programming process keeps a score for each of the dynamic programming paths which is dependent upon the similarity of the parameter frames which are aligned along the path.

In order to reduce the computation required, the dynamic programming algorithm places certain constraints on the way in which the dynamic programming paths can propagate. The constraints employed in this embodiment will now be explained for the case when there are three training utterances. Figure 5 shows a three dimensional cartesian plot, with one dimension provided for each of the three training utterances (Ex1, Ex2 and Ex3). Figure 5 also shows a lattice of eight adjacent points in the three dimensional space which form a cube, with each of the points representing a possible matching between a parameter frame from each of the three training utterances. In this embodiment, the dynamic programming constraints are that if a dynamic programming path ends at point (i,j,k), representing an alignment between the ith parameter frame of the first training utterance, the jth parameter frame of the second training utterance and the kth parameter frame of the third training utterance, then that dynamic programming path can only propagate to the other corners of the cube shown in Figure 5, i.e. to

the points (i+1,j,k), (i,j+1,k), (i,j,k+1), (i+1,j+1,k), (i,j+1,k+1), (i+1,j,k+1) and (i+1,j+1,k+1). When propagating the path to these other points, the dynamic programming process adds the respective "cost" of doing so to the cumulative score for the path ending at point (i,j,k). As those skilled in the art will appreciate, this "cost" depends upon the similarity between the parameter frames represented by the point to which the path propagates. In this embodiment, the seven costs for moving from point (i,j,k) to the seven adjacent points are:

$$C(i+1,j,k)=[d(i+1,j)+d(i+1,k)+d(j,k)]/3+2xPEN$$

$$C(i,j+1,k)=[d(i,j+1)+d(i,k)+d(j+1,k)]/3+2xPEN$$

$$C(i,j,k+1)=[d(i,j)+d(i,k+1)+d(j,k+1)]/3+2xPEN$$

$$C(i+1,j+1,k)=2[d(i+1,j+1)+d(i+1,k)+d(j+1,k)]/3+PEN$$

$$C(i,j+1,k+1)=2[d(i,j+1)+d(i,k+1)+d(j+1,k+1)]/3+PEN$$

$$C(i+1,j,k+1)=2[d(i+1,j)+d(i+1,k+1)+d(j,k+1)]/3+PEN$$

$$C(i+1,j+1,k+1)=[d(i+1,j+1)+d(i+1,k+1)+d(j+1,k+1)]$$

where C(m,n,o) is the cost for moving to point (m,n,o); PEN is a penalty used to discourage too much time compression/expansion of the training utterances during the dynamic programming alignment; and d(m,n) is a similarity score representative of the similarity between parameter frame m and parameter frame n. In this embodiment, d(m,n) is determined using a Euclidean distance measure calculated from the values of the parameters in the parameter frames m and n.

In this embodiment, the dynamic programming process stores the cumulative score for the best dynamic programming path which ends at a lattice point in an associated memory location. In this way, when two or more dynamic programming paths meet during the path propagation, only the path with the best score will propagate further. For example, the score associated with the best dynamic programming path which ends at point (i,j,k) will be stored in the memory location associated with that point. When propagating this path to, for example, point (i,j,k+1), the dynamic programming process determines C(i,j,k+1) and adds this to the cumulative score for the path ending at point (i,j,k) (which is stored in the memory location associated with point (i,j,k)). This updated score is then compared with the cumulative score already stored in the memory location associated with point (i,j,k+1) and is written into this memory location only if the updated score is better than the existing score. A similar procedure is performed for the other six points to which the path ending at point (i,j,k) can propagate.

As those skilled in the art will know, the dynamic programming process begins the propagation of the paths at the start of each of the training utterances, i.e. at the origin of the cartesian coordinates shown in Figure 5 and then propagates the paths using the above

constraints until the paths reach the end of the training utterances. Figure 6 is a flowchart which illustrates the process steps performed by the dynamic programming alignment unit 24 to perform a three dimensional alignment operation. As shown, in step s21, the alignment unit 24 initialises three loop counters (one for each training utterance) i, j and k to zero. Then, in step s21, the alignment unit 24 determines if i is less than the number (NEx1) of parameter frames in the first training utterance, including the noise model at the beginning and end thereof. If it is, then the alignment unit 24 determines, in step s25, if the counter j is less than the number (NEx2) of parameter frames in the second training utterance, including the noise model at the beginning and end thereof. If it is, then the processing proceeds to step s27 where the alignment unit 24 determines if the counter k is less than the number (NEx3) of parameter frames in the third training utterance, including the noise model at the beginning and end thereof. If it is, then the alignment unit 24 propagates, in step s29 the path ending at point (i,j,k) in the manner described above. The alignment unit 24 then increments the counter k in step s31 and the processing returns to step s27.

The processing continues in this way until the counter k has looped through all the parameter frames of the

third training utterance, at which point the processing proceeds to step s33 where the alignment unit 24 resets the counter k and increments the counter j. The processing then proceeds to step s25 and the processing continues in the manner described above until the alignment unit has looped through all the parameter frames in the second training utterance, at which point the processing proceeds to step s35 where the counter j is reset to zero and the counter i is incremented by one. The processing then returns to step s23 where the above procedure is performed again for the incremented value of i. In this way, the alignment unit 24 effectively performs a repeated raster scanning operation of the lattice points, until all the lattice points in the three dimensional space have been propagated in step s29. Once all the points have been processed, the processing then proceeds to step s37, where the alignment unit 24 resets the counter i to zero and determines the optimum alignment between the three training utterances by finding the dynamic programming path with the lowest score. In this embodiment, this is achieved using a standard backtracking algorithm which traces back through path information which is generated and stored for each path during the path propagation process.

Figure 7 illustrates an optimum alignment path which is determined for an example in which the first training

utterance has five parameter frames, the second training utterance has three parameter frames and the third training utterance has four parameter frames. In Figure 7, the best path through the three dimensional grid of lattice points is shown by the bold arrows which are projections of the best path onto the three planes illustrated. For the purpose of this explanation, the origin of the plot shown in Figure 7 is taken to be the alignment of the earliest of the parameter frames in the three training examples (i.e. ignoring any parameter frames of the training utterances which correspond to noise). Labelling this point (0,0,0), then the optimum propagation is as follows:

$(0,0,0) \rightarrow (1,0,1)$

$(1,0,1) \rightarrow (2,1,1)$

$(2,1,1) \rightarrow (3,2,2)$

$(3,2,2) \rightarrow (4,2,3)$

Therefore, the dynamic programming algorithm has found five sets of parameter frames which are aligned with each other. These alignment results are then input to the decoding unit 26 which determines the model for the training utterances.

The above description of the dynamic programming alignment unit 24 was based on the case when there are three training utterances. As those skilled in the art

will appreciate, this technique can be extended to align four, five or any number of training utterances. To do this, the flowchart shown in Figure 6 would be amended to include a loop counter for each of the training utterances and the dynamic programming constraints would have to allow for the propagation of a path to the adjacent parameter frames in each of the training utterances.

DECODING

In this embodiment, the alignment decoding unit 26 generates the word model from the alignment results by combining the parameter frames of the three utterances which are aligned onto an averaged time axis. Therefore, for the example shown in Figure 7, the generated word model would have four parameter frames ([5+4+3]/3). This causes a problem, since the dynamic programming alignment has found five sets of parameter frames which are aligned with each other. Therefore, simply combining the parameter frames in each set would result in a model having five parameter frames. One way of deciding which parameter frames to combine in order to generate the four parameter frames for the model will now be described. For simplicity, it will be described for the case in which there are three training utterances.

In this embodiment, the decoding unit 26 uses the

following function to decide which frames to combine to generate the parameter frames for the model:

$$P = \frac{1}{3}i + \frac{1}{3}j + \frac{1}{3}k + \frac{1}{2} \qquad (1)$$

where i, j and k are the indexes of the aligned parameter frames, i.e. their position in the sequence of parameter frames. It should be noted, that in this decoding step, the parameter frames of the training utterances which are aligned with the noise model 23 are ignored and the first actual alignment between three parameter frames from the three utterances is used as the origin for the indexes i, j, k, as in the example described above with reference to Figure 7. The decoding unit 26 then feeds the alignment results output from the alignment unit 24 into this function and each time its value increases beyond the next integer value, a parameter frame for the model is generated. The way in which this is achieved will be explained for the alignment results discussed above for the alignment shown in Figure 7.

The first set of parameter frames which are aligned correspond to point (0,0,0). Therefore i = j = k = 0, giving p the value of 0.5. The function p is not greater than one, therefore the first parameter frame for the model is not generated. The next set of parameter frames

which are aligned correspond to the point (1,0,1) giving p the value of 1.17. This is greater than one, therefore the decoding unit 26 determines the first parameter frame for the model, but only using the parameter frames corresponding to the point (0,0,0) and not those corresponding to point (1,0,1). It does this by averaging the parameter frames associated with the point (1,0,1). The decoding unit 26 then considers the next set of parameter frames which are aligned, which corresponds to the point (2,1,1). This gives a value of p of 1.83, which is not greater than two and therefore the second parameter frame for the model is not generated. The decoding unit then considers the next set of parameter frames which are aligned, which corresponds to the point (3,2,2). This gives p a value of 2.83, which is greater than two. Therefore, the decoding unit 26 determines the second parameter frame for the model by averaging the parameter frames associated with the points (1,0,1) and (2,1,0) but not those associated with point (3,2,2). The decoding unit 26 then considers the next set of parameter frames which are aligned, which corresponds to the point (4,2,3). This gives p a value of 3.5, which is greater than three. Therefore, the decoding unit 26 determines the third parameter frame for the model by averaging the parameter frames associated with the point (3,2,2) but not using those associated with the point (4,2,3). At this point, there are no more

sets of parameter frames left for the decoding unit 26 to consider and the decoding unit 26 therefore generates the fourth and last parameter frame for the model by averaging the parameter frames associated with the point (4,2,3). As those skilled in the art will appreciate, the above technique will always combine the aligned parameter frames onto an averaged time axis. The resulting model output by the decoding unit 26 can then be used in, for example, a speech recognition system or a speaker verification system.

As those skilled in the art will appreciate, the above decoding technique can be applied to the situation where there are any number of training utterances for which the parameter frames are to be combined on to an averaged time axis. In this case, the generalised function p is as follows:

$$P = \frac{1}{n} \sum_{r=1}^{n} A_r + \frac{1}{2} \qquad (2)$$

where n is the number of training utterances and A is the index of the parameter frame in the training utterance.

As those skilled in the art will appreciate, the above technique is just one way of combining the aligned

parameter frames. An alternative technique is to make the model have the same number of parameter frames as the largest training example. In this case, the decoder would simply combine the parameter frames which are aligned with the parameter frames of the longest training utterance to generate the model. However, this technique is not preferred, since it biases the models towards input utterances which are spoken more slowly.

In the above embodiment, the reference models represented words or phrases input by the user. As those skilled in the art will appreciate, the reference models could be for modelling parts of speech, such as phonemes, syllables or the like.

In the above embodiments, the training utterances were input by the user and the system then generated the reference model from them. As those skilled in the art will appreciate, if the user makes a mistake when entering the utterance or if he utters the wrong command, then the three training utterances will not be consistent. The system may, therefore, perform a consistency check to ensure that the training utterances are consistent with one another. This can be done, for example, using the results of the dynamic programming alignment operation.

In the above embodiment, the dynamic programming process performed a repeated raster scanning operation to propagate the paths through the multidimensional lattice points. As those skilled in the art will appreciate, this is not essential, the dynamic programming paths may be propagated in each dimension at the same rate.

In the above embodiments, a reference model was generated from three or more training utterances. As those skilled in the art will appreciate, the above technique can be used to simultaneously adapt an existing reference model with two or more further training utterances. In such an embodiment, the parameter frames of the existing reference model can be given a weighting relative to the parameter frames of the training utterances which depends upon the number of previous training utterances which were used to generate the existing reference model.

In the above embodiment, the dynamic programming comparison and alignment of the parameter frames in the training utterances was performed without pruning. Pruning is a technique which is often employed in speech recognition systems to reduce the time for performing the dynamic programming matching operation. In this case, the training is performed off-line, and therefore the time involved is not critical. However, if there are a large number of training utterances, then the memory

requirement to perform the dynamic programming process may become very large. In this case, it may be efficient to use pruning to discard badly scoring dynamic programming paths to thereby reduce the memory requirement.

Although the above description relates to a speech recognition system, those skilled in the art will appreciate that the above technique can be used to generate a reference model for use in other applications, such as optical character recognition, handwriting recognition, document template recognition and the like.

CLAIMS:

1. An apparatus for generating a reference model representative of three or more training signals, the apparatus comprising:

means for receiving the three or more training signals;

means for simultaneously comparing and aligning the three or more training signals; and

means for generating said reference model in dependence upon the alignment results from said comparing and aligning means.

2. An apparatus according to claim 1, wherein said means for simultaneously comparing and aligning comprises dynamic programming means for simultaneously performing said comparing and aligning of the three or more training signals.

3. An apparatus according to claim 1 or 2, wherein said combining means is operable to combine the aligned three or more training signals onto an averaged time axis.

4. An apparatus according to any preceding claim, wherein said training signals are representative of speech.

5. An apparatus according to claim 4, wherein said training signals are representative of one or more spoken words.

6. An apparatus according to any preceding claim, wherein one of said training signals is an existing model, and wherein said generating means is operable to place more weight on the existing reference model than the other two training signals.

7. An apparatus according to claim 6, wherein the weighting applied to said reference model by said generating means is dependent upon the number of training examples used to generate the existing reference model.

8. An apparatus for generating a reference model comprising a sequence of reference patterns representative of three or more training signals, the apparatus comprising:

means for receiving the three or more training signals;

means for dividing each training signal into a sequence of time frames and for determining a pattern representative of the signal in each time frame, to generate a respective sequence of patterns for each training signal;

means for simultaneously comparing and aligning the

sequences of patterns representative of the training signals using a dynamic programming alignment technique; and

means for combining the aligned patterns of the three or more training signals to produce said sequence of patterns of said reference model.

9. An apparatus according to claim 8, wherein said combining means is operable to combine said aligned patterns onto an averaged time axis.

10. A method of generating a reference model representative of three or more training signals, the method comprising the steps of:

receiving the three or more training signals;

simultaneously comparing and aligning the three or more training signals; and

generating said reference model in dependence upon the alignment results from said comparing and aligning step.

11. A method according to claim 10, wherein said step of simultaneously comparing and aligning uses a dynamic programming method for simultaneously performing said comparing and aligning of the three or more training signals.

12. A method according to claim 10 or 11, wherein said combining step combines the aligned three or more training signals onto an averaged time axis.

13. A method according to any of claims 10 to 12, wherein said training signals are representative of speech.

14. A method according to claim 13, wherein said training signals are representative of one or more spoken words.

15. A method according to any of claims 10 to 14, wherein one of said training signals is an existing reference model, and wherein said generating step applies a greater weighting to the existing reference model than to the other training signals.

16. A method according to claim 15, wherein the weighting applied to said existing reference model depends upon the number of training examples used to generate the existing reference model.

17. A method of generating a reference model comprising a sequence of reference patterns representative of three or more training signals, the method comprising the steps of:

receiving the three or more training signals;

dividing each training signal into a sequence of time frames and for determining a pattern representative of the signal in each time frame, to generate a respective sequence of patterns for each training signal;

simultaneously comparing and aligning the sequences of patterns representative of the training signals using a dynamic programming alignment technique; and

combining the aligned patterns of the three or more training signals to produce said sequence of patterns of said reference model.

18. A method according to claim 17, wherein said combining step combines said aligned patterns onto an averaged time axis.

19. A computer readable medium carrying instructions for configuring a programmable processor to be configured as a reference model generating apparatus according to any of claims 1 to 9.

20. A signal carrying instructions for configuring a programmable processing circuit as a reference model generating apparatus according to any of claims 1 to 9.

21. A computer readable medium storing computer executable process steps for generating a reference model

representative of three or more training signals, the process steps comprising:

steps for receiving the three or more training signals;

steps for simultaneously comparing and aligning the three or more training signals; and

steps for generating said reference model in dependence upon the alignment results from said comparing and aligning steps.

22. Computer executable process steps for generating a reference model representative of three or more training signals, the process steps comprising:

steps for receiving the three or more training signals;

steps for simultaneously comparing and aligning the three or more training signals; and

steps for generating said reference model in dependence upon the alignment results from said comparing and aligning steps.

Application No:	GB 9909429.4	**Examiner:**	John Donaldson
Claims searched:	1 to 22	**Date of search:**	15 May 2000

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.R): G4R(RPD, RPN, RRL)

Int Cl (Ed.7): G06K 9/00, 9/20, 9/32, 9/60, 9/62, 9/64, 9/66, 9/78, 9/80; G10L 15/00 15/06, 17/00

Other: Online:WPI, EPODOC, JAPIO

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
A	US 4751737 (GERSON), see abstract	-

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.

CID: <GB__2349260A__I_>